www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issues 6 June 2016, Page No. 16875-16879

Survey on Pre-Processing Techniques for Text Mining

Arjun Srinivas Nayak[1], Ananthu P Kanive[2], Naveen Chandavekar[3], Dr. Balasubramani R[4]

[1,2]NMAM Institute of Technology,
Dept. of Computer Science and Engineering
[1]arjunsnayak12@gmail.com, [2]ananthukanive@live.com

[3]NMAM Institute of Technology,
Dept. of Computer Science and Engineering
Chandavarkar@nitte.edu.in

[4]NMAM Institute of Technology,
Dept. of Information Science and Engineering
balasubramani.r@nitte.edu.in

Abstract: ***Data Mining is a versatile sublet in the field of computer science. It is the computational evolution mode of detecting patterns in large data sets. This paper give an indication on the different pre-processing techniques to mine text data. Text mining applications include – Information Retrieval, Information Extraction, Categorization, and Natural Language Processing. The pre-processing of text mining starts with Tokenization, followed by Stop-word removal and finally stemming. This paper evaluates Porter's and Krovetz algorithm, highlighting their applications and drawbacks.***

Keywords: Tokenization, Stop-word Removal, Stemming, Porter, Krovetz.

1. INTRODUCTION

Data pre-processing is an often neglected but important step in the data mining process. Pre-processing involves techniques to transform raw data into more understandable format.

Data gathered from real world instances is usually incomplete, noisy and inconsistent. Incomplete data lacks attribute values, lacks certain attributes of interest, thereby containing only aggregate module. Noisy data comprises of errors and outliers. Data that is inconsistent contains discrepancies in codes or names.

Pre-processing mainly involves:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Data Discretization

Data Cleaning converts the raw data by filling in missing values, smoothens noisy data, identifies errors and removes outliers, and resolves the inconsistencies.

Data Integration combines data from several multiple sources into one coherent data store, using multiple databases, data-cubes or files. Data Transformation involves two steps: Normalization fits the data within a range, & Aggregation combines the normalized data. Data Reduction reduces the volume of the data producing the same end result. Data Discretization is a part of data reduction that replaces the numerical attributes with nominal ones.

The data/information present in documents, e-mails and other data files are generally categorized as structured, semi-structured and unstructured data.

Structured Data is well organized, follows a consistent order

and can be readily accessed and read by a person or a computer program.

This type of data is stored in well-defined schemes such as data-bases. Excel spreadsheets, library catalogues, phone book and statistical tables are some of the examples.

Unstructured data, on the other hand, is the complete opposite of structured data. It tends to be free form, non-tabular, dispersed, and not easily retrievable. Hence, it requires deliberate intervention to make sense of it. E.g. Emails, Web Pages, Files like text, audio and video.

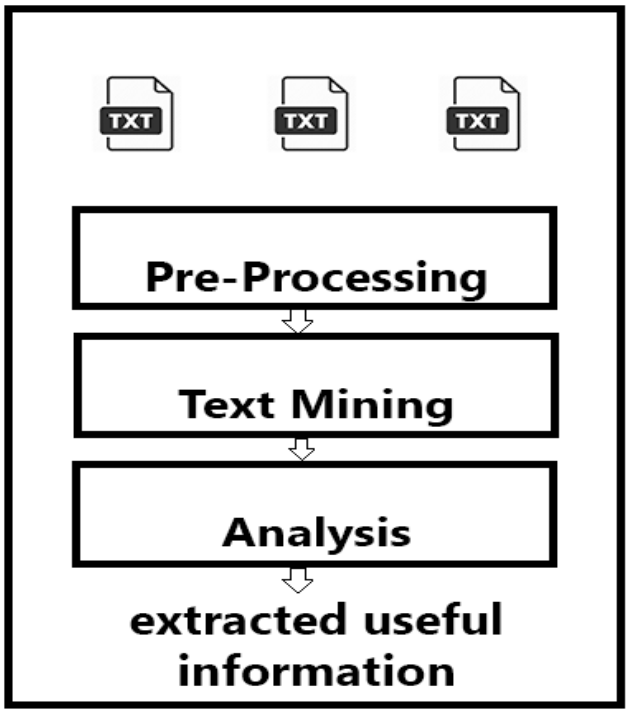
Due to its complexity, it is hard to categorize its contents. Unstructured data is usually text, created in free-form styles and finding any attribute to describe it is a tedious task. Computer programs cannot analyse or generate reports on such data simply because it lacks structure, it has no underlying dominant characteristics or the individual items of data have no common ground.

In general, a higher percentage of data is found to be unstructured. Since it's difficult to mine this data, it requires more attention and hence consumes higher processing time. Thus, pre-processing improves the efficiency of the mining algorithms on unstructured data.

In text mining, pre-processing is a 3 step mechanism composed of Tokenization, Stop Word Removal and Stemming.

Stemmers are used to consolidate terms to optimize retrieval performance and/or to reduce the size of indexing files. Stemming will, in general, increase memory at the cost of decreased precision. Studies of the effects of stemming on retrieval effectiveness are equivocal, but generally stemming has either no effect, or a positive effect, on retrieval performance where the measures used include both memory and precision.

Several Stemming Algorithms have been developed over the years to optimize the data. Porter's Algorithm is one of the efficient techniques for the English Language. Krovetz Algorithm proves to be better for light English text. But each have their own pros and cons.



Text Mining Process

2. LITERATURE SURVEY

Dr. S Vijayarani[2] discussed the purpose of pre-processing of text data, highlighting the applications of text mining and its various contingencies. Text mining is the process of seeking or extracting the useful information from the textual data. It tries to find interesting patterns from large databases.

Information Retrieval is the association and retrieval of information from a large number of text based documents. Information Extraction identifies keywords and relationships within the text, by using pattern matching techniques and converting it to a relational database.

Categorization involves identifying the main themes of a document by comparing the document to a pre-defined set of topics. Unlike Information Extraction, it doesn't try to process the actual information.

Natural Language Processing tells how computers can be used to understand and manipulate the natural language text. Applications of NLP include a number of fields of study, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence and expert systems etc.

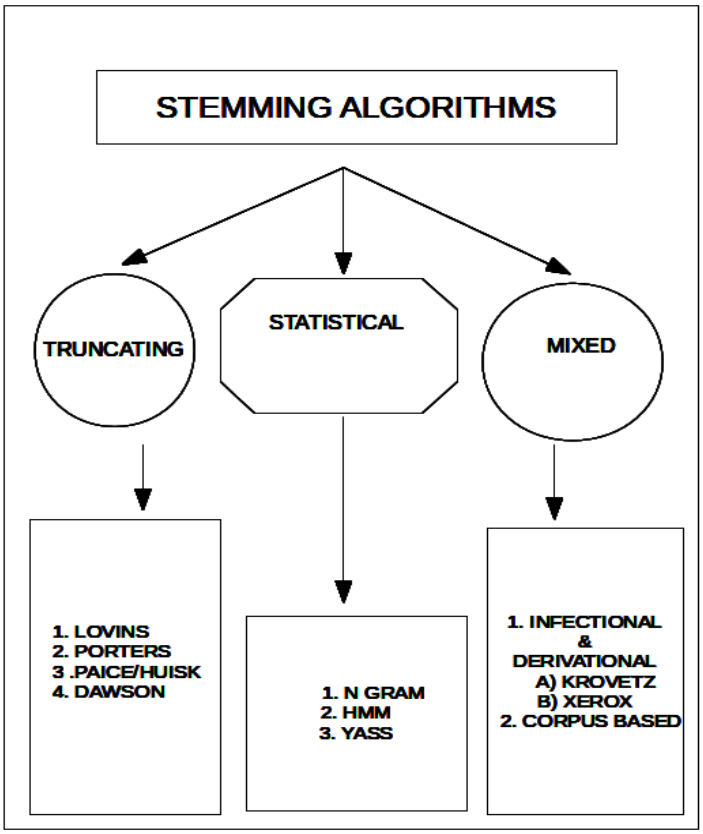
Pre-processing mechanism include **Extraction, Stop words Elimination and Stemming.**

Extraction method is used to tokenize the file content into individual elements.

Stop Words Elimination aims to make the text look heavier and less important for analysis by removing stop words to reduce the dimensionality of term space.

Stemming finds the root or stem of the words that are phonologically related, i.e., removing the common suffixes, reducing the number of words, to accurately match stems.

Various stemming algorithms have been developed over the years, each having its own purpose, providing services for different domains.



"Affix Removal ".

The study revolves around pre-processing techniques that eliminate the noise from the text data, performs stemming, and trims the size of the text data.

Vairaprakash Gurusamy et.al [1] analysed the importance of pre-processing in text mining, natural language processing and information retrieval. They have evaluated the issues in pre-processing methods for text archives.

The paper examines the need for text pre-processing in NLP systems. Their proposed pre-processing methods use **Tokenization, Stop Word Removal, and Stemming**.

Tokenization probes the sentences and makes a list of tokens which can be used as input for further algorithms. The predicament include removal of elements such as brackets, hyphens and other punctuations. In addition to that, tokenizing languages other than English can be laborious.

Stemming revolves about finding the common representation of words. The flaws in this process are **over-stemming and under-stemming.**

Mainly discussed stemming methods are "Table Look Up Approach", "Successor Variety ", "n- Gram stemmers" and

C.Ramasubramanian et.al [3] from ANNA University worked on ways to improve the stemming techniques used in the pre-processing of text mining. The paper looks into the disadvantages of one of the stemming algorithms called MF Porter's algorithm. Drawbacks of the existing approach is discussed along with the process to overcome them.

Spell check is added to overcome wrong matches and increase the accuracy. This saves processing time for the misspelled words. The proposed a Smart Word List that removes Stop Words efficiently without removing the important words.

Vikram Singh et.al [6] from National Institute of Technology, Haryana, have worked on some effective pre-processing techniques for Information Retrieval Systems. The paper focuses on Tokenization and stemming algorithms that can be used when pre-processing data.

Information Retrieval systems see text data as a bag of words. Its responsibility is to give a "representation that indicates what

the document is about and what topics it (the document) covers".

The paper views tokenization as one of the crucial steps in pre-processing text data. It splits this "bag of words" into identifiable words known as tokens. Tokenization also gives information indicating the frequencies of each token, which can be used in further steps of Information Retrieval. Stop-word removal and Stemming algorithms are then applied and the output will contain only those tokens that are deemed valuable by the pre-processing algorithms.

With some sample data sets, an experiment shows up to 68% increase in efficiency, in reference to the tokens generated after pre-processing compared to that generated without pre-processing.

Cristian Moral et.al [7] from Universidad Politécnica de Madrid, Spain have conducted an assessment of stemming algorithms used in Information Retrieval applications.

A stemmer has three main purposes:

- Clustering words according to topic
- Improves the effectiveness of IR algorithms
- Words sharing the same stem leads to reduction of the number of words that needs to be processed

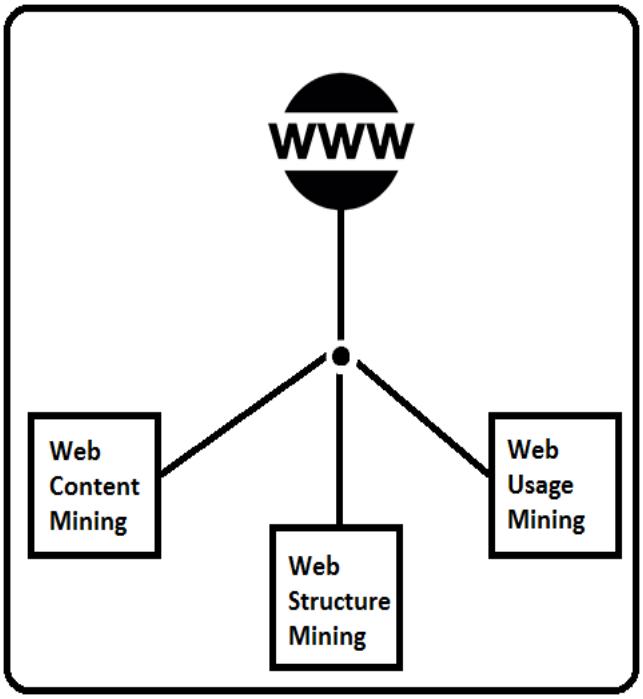
Porter's Stemmer (1980) is the most used algorithm in information retrieval applications, although the Lovins stemmer is considered to be the most widely known algorithm. Porter's algorithm is a five step process, which is applied to every word in the input data set. According to the algorithm, a word is defined as $[C](VC)^m[V]$ where C and V are lists of consonants and vowels respectively and m is the measure of the word. It uses about sixty rules, divided into 5 groups to accurately derive the stem of a word.

Another stemming algorithm called Krovetz's Algorithm, is a simple stemmer that covers plurals, past tense and –ing verb forms. It supports a dictionary which can be used to verify the stemmed words.

The application of stemmers in information retrieval applications may not be widely agreed upon, but experiments demonstrate that the nature of the document has a large impact on the output generated by these pre-processing algorithms.

Parth Suthar et.al [4] from L.D College of Engineering, Ahmedabad, examined on Web Usage Mining Techniques.The survey paper reviews Web Content Mining. The content on the Web can be unstructured, semi- structured or structured.

It can be located in web servers, databases or HTML pages. Various techniques are employed to mine this data to find usual information. This data can be used to find surfing habits



Web Mining

of users and their patterns of interests.

The first step is pre-processing the raw usage data. This raw data is usually very substantial and noisy.

The steps involved are Data Cleaning, User Identification, Session Identification and Path Completion.

3. CONCLUSION

In this paper, we have surveyed various pre-processing techniques for mining unstructured data and its applications. Text

We've looked into the Stemming algorithms such as MF Porter and Krovetz and have analysed their efficiencies and accuracies. Some of the drawbacks of Porter's algorithm are that it leads to a large degree, it is context dependant, results into a wrong stem etc.

The major drawback of Krovetz stemming algorithm is that it becomes inefficient with large input documents and its inability to deal with words that are not present in the lexicon.

There's scope for research in this area to improve the efficiency of these two algorithms to enhance the Stemming process.

REFERENCES

[1] Vairaprakash Gurusamy, Subbu Kannan: *"Preprocessing Techniques for Text Mining"* October 2014

[2] Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya: *"Preprocessing Techniques for Text Mining - An Overview"* International Journal of Computer Science & Communication Networks, Vol 5(1),7-16

[3] C.Ramasubramanian, R.Ramya: *"Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm"* International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013

[4] Parth Suthar, Prof. Bhavesh Oza: *"A Survey of Web Usage Mining Techniques"* International Journal of Computer Science and Information Technologies, Vol. 6 (6), 2015, 5073-5076

[5] "*Data Preprocessing Techniques for Data Mining",* Winter School On "Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets

[6] Vikram Singh and Balwinder Saini "An Effective Pre-Processing Algorithm For Information Retrieval Systems" International Journal of Database Management Systems (IJDMS) Vol.6, No.6, December 2014

[7] Moral, C., de Antonio, A., Imbert, R. & Ramírez, J. (2014). A survey of stemming algorithms in information retrieval/ Information Research, 19(1) paper 605. [Available at http://InformationR.net/ir/191/paper605.html]

[8] **Author Profile**

Mr. Arjun Nayak is pursuing his Bachelors of Engineering in Computer Science and Engineering at NMAMIT, Nitte (an autonomous institution under VTU Belgaum). His major work is in the area of Data Mining and Machine Learning. His areas interests are programming and web designing.

Mr. Ananthu P Kanive is pursuing his Bachelors of Engineering in Computer Science and Engineering at NMAMIT, Nitte (an autonomous institution under VTU Belgaum). His major work is in the area of Data Mining and Software Automation. His areas interests are programming and game development.

Mr. Naveen Chandavarkar obtained his B.E in Computer Science and Engineering from VTU in 2008 and M.Tech in Computer Science and Engineering in 2012. He is pursuing his Ph.D in Computer Science and Engineering from VTU in the field of Opinion Mining. He is having 6 years of teaching experience. He is a life member of Indian Society of Technical education. He has several papers published in the area of Data Mining.

Dr. Balasubramani obtained his BE in Electronics and Communication from Madurai Kamaraj University in 1990 and M.Tech in Information Technology from AAI-Deemed University, Allahabad in 2005. He obtained his Ph.D., in Information Technology from Vinayaka Missions University, Salem in 2011 for his research in the area of Digital Image Processing. He is having 26 years of professional experience (12 years in industry & 14 years in teaching). Presently he is working as the Professor & Head in the Dept. of ISE at NMAMIT, Nitte.